科研人员职业高峰前后的研究主题转换特征识别*

■ 陈立雪 滕广青 吕晶 庹锐
东北师范大学信息科学与技术学院 长春 130117

摘 要：[目的/意义]探索科研人员职业发展情况及其研究主题的变化规律不仅可以揭示科学生产力发展的内在机制，也有助于对科学事业的发展提供更好的政策指导与支持。[方法/过程]基于自然科学、社会科学、艺术与人文科学的代表性学科数据，对科研人员的职业高峰进行识别。在此基础上以职业高峰作为科研人员学术生涯的划分依据，采用自然语言处理中的Top2Vec主题建模方法识别研究主题，对科研人员学术生涯不同阶段所研究主题的主题相似度和主题转换概率进行分析。[结果/结论]研究结果表明，各学科科研人员总体上在经历职业高峰之后的主题转换会更加频繁；而精英学者在经历职业高峰后其研究主题则反而更加专一。

关键词：科研人员 职业高峰 Top2Vec 主题转换 主题相似度
分类号：G250.2
DOI：10.13266/j.issn.0252-3116.2021.16.009

1 引言

对于科研人员的职业生涯变化规律及其主题变迁的研究一直以来都是图书情报学领域的研究热点，尤其是对科研拔尖人才的研究更是社会与学术界关注的重点[1]。根据马太效应[2]，科学家个体在职业生涯中取得优秀成绩能够带来声誉和认可。这些声誉和认可往往可以转化为有形资产，反过来有助于其未来职业的成功。最近发表在Nature上的一项研究也发现科研人员职业生涯中通常会涉及一段“高光时期”(hot streak)。在这段时期内科学家个体的表现会大大高于其正常表现，最为显著的特征就是科学家个人在这段时期内的成果备受瞩目(科研成果被高频引用)[3]。尽管现有的研究发现在科学家职业生涯中存在类似的高光时期或高峰期，但是鲜有研究去深入挖掘职业高峰前后科研人员个体的科研工作到底发生了何种变化，特别是科研人员以及精英学者们在职业高峰期前后其研究主题发生了怎样的变化。2019年6月，中共中央办公厅和国务院办公厅在《关于进一步弘扬科学家精神加强作风和学风建设的意见》[4]中指出：“要加大对优秀科技工作者和创新团队的稳定支持力度，以加快培育促进科技事业健康发展”。从这个角度来说，对科研人员尤其是优秀科研人员活动机制的研究也是为了对科学事业的进一步发展提供更好的政策指导与支持。因此，有必要在实施国家科技发展战略的大环境下，对科研人员尤其是优秀科研工作者的科研学研究活动的特征进行细致地探索与分析。

由于知识的发展是连续的、流动的和多领域交叉的，科研人员所研究主题的变化反映了信息收集与知识传递的不断变化[5]。另外，近年来科学知识迅猛发展，新问题、新知识层出不穷。有鉴于此，笔者尝试结合科研人员职业高峰与研究主题两个维度，分别从自然科学、社会科学、艺术与人文科学中选择不同学科领域的数据，采用自然语言处理(NLP)方法，从科研人员职业高峰的视角对科研人员所研究主题的变化进行具体分析，以期对科研人员以及精英学者们在职业高峰前后研究主题的变化特征取得更清晰的认知和更深入的洞见。

2 相关研究现状

了解科学家个体研究活动机制及其学术生涯过程中的重要里程碑，有助于深入探索科学生产力的动态

* 本文系国家社会科学基金项目“基于复合数据的科技信息跨维度挖掘与推荐研究”(项目编号：19BTQ063)研究成果之一。
作者简介：陈立雪(ORCID：0000-0002-4661-4679)，硕士研究生；滕广青(ORCID：0000-0002-1053-0959)，教授，博士生导师，通讯作者，E-mail：tengguangqing@163.com；吕晶(ORCID：0000-0003-2482-5827)，硕士研究生；庹锐(ORCID：0000-0002-7207-8750)，博士研究生。
收稿日期：2021-04-11 修回日期：2021-07-18 本文起止页码：81-89 本文责任编辑：徐健

模式。从社会学理论来讲,年轻科学家作为学术界的“边缘人”,在特定想法或学术流派中的投入尚少,没有积累较多的声誉,因此不用过分担心科研失败带来的损失,往往也更容易做出成绩,同时年轻科学家善于从新视角去看待老问题,他们兴趣更为广泛、精力更加充沛、学术热情更高,尽管他们缺乏经验,但研究原创性高;年老科学家虽在研究经验的积累、独立判断、处理矛盾等方面更胜一筹,但他们缺少热情,会产生许多没有灵感的作品也就不容易做出重大突破[6,7]。B. F. Jones 等[8]通过对诺贝尔奖学者的职业生涯研究,发现富有想法的年轻人更容易在硬科学(hard science)研究中做出重大突破。此外,学术界有许多研究工作对科研人员的职业高峰及其所对应的科研成就展开了研究[9-12]。这些研究工作虽然对科研人员的学术生涯发展给予了高度重视,但是对职业高峰的界定并不统一,研究视角也相对单一,并没有关注伴随科研人员职业高峰的科研工作发生了怎样的变化。在 2020 年最新的一项研究中,研究者在证实诺贝尔奖得主比其他科学家在学术生涯早期就拥有更多的发文量与更高的被引量的同时,还发现了获奖后得主们科研成果影响力下滑的短暂的“诺贝尔低谷”(Nobel Dip)现象[13]。这意味着科研人员在经历了职业高峰之后,在具体的科学工作中会发生一些有趣的变化。其中,科研人员职业高峰前后研究主题的变化成为学术界关注的一个问题。

具有前瞻性的主题可能会促使高影响力研究成果的产生,这不仅可以提高科学家的声誉,也可以给整个领域创造研究机会。鉴于研究主题对科研人员个体学术生涯以及对学科和创新政策的影响,迫切地需要采取定量方法来理解科学家们在整个学术生涯中其研究主题是如何变化的[14-16]。近年来,国内外学术界均有学者聚焦于量化和模拟科学家学术生涯中研究主题的演变[17-20]。尽管研究主题的频繁变化可能会带来失败和生产力下降的风险,但是也有研究表明一个稳定而又有重点的研究团队虽然有助于科学家保持生产力,但却不利于创新[21,22]。通常而言,科研人员在其学术生涯过程中所研究的主题内容不可能是一成不变的,科学家转换自己的研究主题可能是在保守与冒险之间权衡的结果[23]。A. Hoonlor 等[24]选择计算机领域的期刊与会议论文进行分析发现,科学家的研究重点大约以 10 年为一个周期发生变化,只有少部分研究者在同一主题年复一年地长期发表文章;A. Rzhetsky 等[25]将学科知识建模为网络,通过分析发表在 30 多年内的数百万篇生物医学论文发现,生物医学领域的科学家越来越追求保守的研究策略,倾向于探索中心主题的局部邻域而不是进行大跨度的主题转换;T. Jia 等[26]则以物理学领域的分类代码为依据,发现物理学家的研究兴趣从学术生涯的开始到学术生涯的结束,其间发生了极大的转变;A. Zeng 等[27]在最近的一项研究中发现,如今的科研人员相比更早的研究者更频繁地在不同主题之间切换,并且学术生涯早期的高转换率与较低的整体生产力有关。

综上,学术界关于科研人员职业高峰与学术生涯中研究主题转换的相关研究已经分别积累了一定的成果。但在现有的研究中,鲜有学者将科研人员个体职业高峰与其研究主题转换联系起来进行分析。有鉴于此,笔者从自然科学、社会科学、艺术与人文科学 3 个学科领域中分别选取代表性学科,对科研人员以及精英学者们学术生涯中不同阶段的研究主题变化特征进行深入研究,以期为揭示科学生产力发展机制提供可资借鉴的参考。

3 相关理论基础

3.1 科研人员研究主题识别

识别科研人员的研究主题,主要是通过对其已发表的成果文献进行自然语言处理(NLP),从中发现大型文档集合中的潜在语义结构,通常也被称为主题分类。当前应用最广泛的主题建模方法包括概率潜在语义分析(probabilistic latent semantic analysis, PLSA)[28]和隐含狄利克雷分布(latent dirichlet allocation, LDA)[29]等方法。尽管这些建模方法在学术研究中很受欢迎,但也存在一些缺陷。比如为了使模型达到最佳效果,在建模之前通常需要做一些预处理,如自定义停用词列表、进行词干提取、词元化以及花费大量精力去预先设置合适的主题数量等。此外,大部分主题建模方法依赖于文档的词袋表示,忽略了单词的顺序和语义。为了克服这些缺陷,笔者采用 2020 年最新提出的 Top2Vec[30]主题建模方法对科研人员公开发表的文献进行主题建模以识别其研究主题。

Top2Vec 作为一种分布式主题向量模型,它利用文档和单词的语义嵌入来寻找主题。在语义空间中发现的文档密集区域的数量被认为是突出主题的数量。其中,主题向量是从文档的密集区域中计算出来的,密集区域是由非常相似的文档组成的,通过计算“质心”(centroid)来得到主题向量,即同一密集簇中所有文档向量的算术平均值。“质心”能够很好地代表文档密

集区域的主题向量,最接近这个主题向量的词也就是在语义上能够最好地描述它的词。最终得到的主题向量与文档和词向量的共同嵌入,词向量之间的距离表示语义相似度。Top2Vec 生成的主题也被证明比概率生成模型具有更大的信息量以及包含更具有代表性的语料。该模型不需要去停用词,也无需进行词干提取和词元化等预处理,它可以自动查找主题数量。其主要操作过程如图 1 所示：

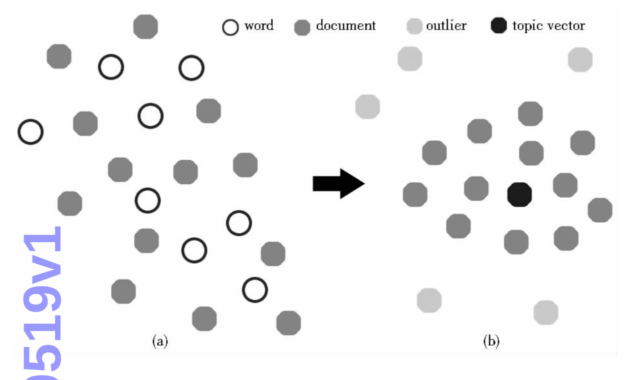


图 1　基于 top2vec 主题建模的操作过程示例

在图 1 中,Top2Vec 主题建模的具体操作步骤如下:①创建语义嵌入。使用 Sentence Transformer 创建嵌入的文档和词向量。图 1(a)显示了一个语义空间的示例。灰色的点是文档,空心的点是单词。单词最接近它们最能代表的文档,相似的文档也靠得很近。②使用 UMAP(uniform manifold approximation and projection)创建文档向量的低维嵌入。高维空间中的文档向量非常稀疏,降维有助于发现密集区域,其中每一点都是一个文档向量。③使用 HDBSCAN 查找文档的密集区域并计算主题向量。对文档向量采用基于层次密度的空间聚类(HDBSCAN),以聚类数量代替主题数量。HDBSCAN 作为一种基于密度的聚类方法对于所识别的异常值不用于计算“质心”,它不会强制每个文档都必须分到某一类别,而是将这些未被分入主题集群的文档设为离群值。在图 1(b)中,浅灰色点即是不属于特定集群的异常值。对每一组属于密集聚类的文档向量进行“质心”计算,为每一主题生成一个主题向量(图 1(b)中黑色点)。④基于 C-TF-IDF 的关键词提取。完成主题聚类之后还需要基于内容探究一个主题与另一个主题的不同。采用基于 TF-IDF[31] 的变体 C-TF-IDF 进行主题词探索,C-TF-IDF 是一个基于类的 TF-IDF 过程,其中 C 表示 CLASS 类,它可以根据文本文档所在的主题类别提取它们的生成特性。与传统的 TF-IDF 不同的是,C-TF-IDF 并非用来比较不同文档之间单词的重要性,而是将单个主题类别中所有文档作为单个文档处理,每个类别可以被看作是一个非常长的文档,所得到的 C-TF-IDF 分数可以反映一个主题中重要单词的权重。它可以提取使每个主题有相对于其他主题独特的元素。公式(1)如下所示：

$$C-TF-IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j} \qquad 公式(1)$$

在公式(1)中,对每一类主题 i 提取每个单词 t 的频次 t_i,除以该主题总单词数 w_i,这是主题高频词汇的一种规则化形式;再用文档总数 m 除以单词 t 在所有类 n 中的总出现频次,将其转化为对数形式后与前一项相乘,以此完成科研人员研究主题的识别。

3.2　主题相似性和主题转换概率

笔者选取主题相似度得分和主题转换概率两个指标来测量科研人员研究主题的变换情况。相似度得分可以衡量科研人员在不同主题转换过程中到底进行了多大幅度的主题迁移;主题转换概率用于判断科研人员研究主题转换频率的高低。研究工作采用余弦相似度计算主题间的相似性得分,该方法已经被证实是当前自然语言处理中应用最广泛的语义距离测度方法。余弦相似度(Cosine Similarity)算法是根据两个词向量之间的余弦夹角判断词向量之间的相似性,余弦值越接近 1,就表明夹角越接近 0 度,也就是两个向量越相似,夹角等于 0,即两个向量相等,公式(2)如下所示：

$$\text{Similarity} = \cos(\theta) \frac{A \cdot B}{\| A \| \ \| B \|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2 \times (B_i)^2}} \qquad 公式(2)$$

其中,A_i 和 B_i 表示的是两类主题所包含的文档词向量的特征,$\| A \|$ 和 $\| B \|$ 是两个词向量的 L2 范数,θ 是词向量 A 和 B 之间的角度,Similarity 表示余弦相似度的最终得分。基于已经完成的主题识别,采用以上公式计算不同主题之间的相似性。根据主题相似度得分分别构建单个科研人员研究主题相似度得分矩阵,如表 1 所示：

表 1　单个科研人员主题相似度得分矩阵

	w_1	w_2	w_3	w_4	…	w_n
w_1	s_{11}	s_{12}	s_{13}	s_{14}	…	s_{1n}
w_2		s_{22}	s_{23}	s_{24}	…	s_{2n}
w_3			s_{33}	s_{34}	…	s_{3n}
w_4				s_{44}	…	s_{4n}
…	…	…	…	…	…	…
w_n				…	s_{nn}	

在表 1 中,n 表示文献的总数,w_n 表示科研人员的第 n 篇文献,s_{ij} 表示依据余弦相似度得出的文献 i 和文

献 j 的相似度分数。在此基础上，采用公式(3)计算单个科研人员某段时期内所有文献的主题相似度得分。

$$SIM_{au}=\frac{2\sum_{ij=1}^{ij=n}s_{ij}}{n(n+1)}\quad(1\leqslant i\leqslant j\leqslant n)\qquad 公式(3)$$

在公式(3)中，SIM_{au} 表示单个科研人员某段时期内所有文献的相似度得分。一段时间内某科研人员所发表论文主题的相似度得分越低，说明在这段时期内科研人员研究主题的跨度越大。此外，个体科研人员的主题转换概率计算公式如下所示：

$$Switch\ Probability=\begin{cases}1, & t_n=n\ and\ n\neq 1\\ 0, & t_n=n=1\\ \frac{t_n-1}{n-1}, & t_n\neq n\end{cases}$$

公式(4)

其中，n 表示科研人员发表论文的总数，t_n 表示科研人员所包含文献的主题数量。*Switch Probability* 表示主题转换概率，主题转换概率越高表示科研人员在不同研究主题之间转换的概率越频繁；主题转换概率越低，表示该科研人员的研究主题越专一。

4 研究方法与流程

4.1 数据来源与流程框架

在多学科视角下(自然科学、社会科学、艺术与人文科学)探测科研人员高峰期前/后的科研主题变化特征，需要在以往仅针对某单一学科领域的基础上考虑更多的因素。针对单一学科的研究无需考虑文档数量因素，但多学科视野下学科间差别悬殊的文档数量可能会给主题建模与统计结果造成偏倚，不利于学科间的横向比较。基于这一原因，笔者选取了真菌学、图书情报学、哲学 3 个在文档数量上大体相当的学科分别作为自然科学、社会科学、艺术与人文科学的代表。以 Web of Science 核心数据库作为基础数据来源，采用高级检索，检索式分别为“SU = MYCOLOGY”“SU = INFORMATION SCIENCE & LIBRARY SCIENCE”“SU = PHILOSOPHY”，检索日期为 2020 年 11 月 1 日，检索时间段为 1985 年至今，将文献类别限定为“Article”，语种限定为“English”，最终获得 158 446 篇文献。其中，真菌学文献 43 000 篇，图书情报学文献 65 961 篇，哲学文献 49 485 篇。在此基础上，进一步提取文献中所包含的作者，并按照所属学科进行分组。根据 ORCID 标识符对重名作者进行人工核查且不重复计数，共得到 266 388 位作者。其中，真菌学 113 241 位，图书情报学 106 730 位，哲学 46 417 位。

此外，考虑到原 Top2Vec 算法所依赖的 TensorFlow-Text 安装包对 Windows 系统的限制，因此为了使研究方法具有更好的泛化性和研究复现性，笔者在深度学习 PyTorch 框架下使用基于 Top2Vec 的主题建模方法。相比于原本的 Top2Vec 建模方法，不仅保留了原模型的内核，同时具有更好的系统兼容性。

传统认知下，重要奖项、高水平成果等都可以作为科研人员职业高峰的标志。但学术界中重要奖项凤毛麟角，不足以据此考量更广泛的科研人员队伍。且奖项更侧重学术界对此前成就的认可，而不是科学研究本身在获奖时达到最高峰。因此，学术界主要采用高被引论文作为识别科研人员职业高峰的依据，特别是针对长时间周期某单一学科进行研究时，主要采用设定统一年限(如 10 年)内的引文频次[3, 32]。考虑到本研究对象跨越 3 个学科门类，且论文半衰期(half life)受到文献类型、学科性质等多方面因素的影响，不适合采用统一年限内的引文频次作为标准，加之“睡美人”文献等因素的影响，笔者使用绝对被引频次最高的论文作为科研人员职业高峰的标志，并将最高被引发表当年视为该科研人员达到职业高峰。

具体的研究工作按照以下流程：①采用 Top2Vec 模型对所获得文献进行主题建模，并对小众主题进行归并；②计算高峰期前/后各自时段内部主题相似度与主题转换概率，比较分析科研人员整体在高峰期前/后各时段内部的主题转换特征；③筛选精英学者，计算精英学者高峰期前与高峰期后的主题相似度与主题转换概率差值，分析精英学者经历高峰期前后的主题转换特征。

4.2 主题分布总体概况

研究工作将真菌学(43 000 篇)、图书情报学(65 961篇)、哲学(49 485 篇)文献的摘要作为研究的基础语料数据，采用基于 Top2Vec 的主题识别方法对其进行主题建模。建模过程中预先设置阈值，限制每个主题集群所包含的文档数量不少于 20 篇。分别得到 3 个学科 1985 年至今研究主题的分布结果见图 2。

在图 2 中，(a)、(b)、(c)分别为真菌学、图书情报学、哲学 3 个学科研究主题分布的可视化结果。图中的节点为已发表的文献，不同深浅程度的区域所表示的是不同的主题，浅灰色表示的是不被归为任意一个主题的离群值，对突出主题的高权重主题词进行了标记，这些主题词代表了其所属主题的关键信息。

真菌学经过主题建模得到 56 个主题。从图 2(a)中可以清晰地看出该学科研究主题分散度非常高（充

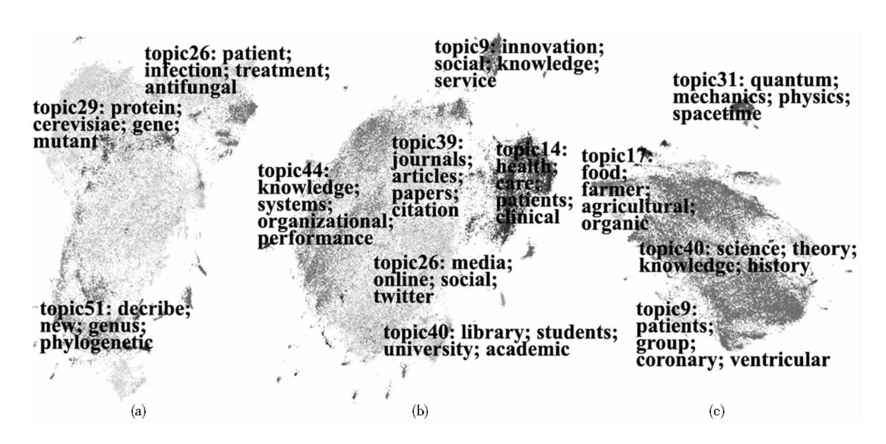


图2　各学科主题分布

斥大量浅灰色离群值），缺少聚集海量文献的主题，且离群值（文献）数量较多。包含文献数相对较多的主题有topic26、topic29和topic51。其中，patient（病人）、infection（感染）、treatment（治疗）和antifungal（抗菌的）这些高权重主题词说明topic26主要涉及的内容是传染病研究；protein（蛋白质）、cerevisiae（菌株）、gene（基因）和mutant（突变）这些高权重主题词则反映topic29是基因遗传有关的研究；topic51中的高权重主题词有decribe（描述）、new（新的）、genus（物种）和phylogenetic（种系发生的），表明该主题是生物物种相关研究。除此之外，其他的主题所包含文档数量都较少，但是该学科研究主题总数又很多，这也在一定程度上显示了真菌学学科的复杂与多样性。

图书情报学的主题建模结果得到了46个主题。在图2（b）中可以看到，最为突出的主题是topic14、topic44、topic26、topic39、topic40和topic9。根据各主题中的高权重主题词可以推断出各主题的研究分别集中在医学信息学（topic14）、知识组织（topic44）、网络信息传播（topic26）、文献计量（topic39）、高校图书馆（topic40）、知识服务（topic9）领域。这六大主题构成了图书情报学学科的核心研究主题。除了核心主题，该学科还存在其他包含文档数量较少的小众主题，并且这些小众主题大多游离在中心主题外的边缘区域，这也说明小众主题与核心主题的研究内容差别较大。

哲学学科的主题建模结果得到41个主题，如图2（c）所示。可视化结果显示该学科的主题具有高度的聚集性，绝大多数文献被归入位于中央区域的topic40。该主题的高权重主题词包括science（科学）、theory（理论）、knowledge（知识）和history（历史），属于哲学基础理论研究。此外，较为突出的主题还有topic9、topic17和topic31。对照各个主题中的高权重主题词不难发现，其研究内容主要为医药哲学、农业哲学、科学哲学等，这也是此类主题分布处于大量文献集群之外的边缘区域的一个原因，这类主题不与其他任何一个主题有高度的主题相关性。

综合上述情况可以发现，哲学主题对文献的聚集性最高且学科内部主题数量最少；真菌学主题分布最为松散且主题总数也最多；图书情报学介于二者之间，但其聚类结果中包含大量文献的主题数量是最多的。考虑到本研究基于Top2Vec主题建模方法所得到的主题分布结果中，各学科的众多主题中包含大量基于少量文献的小众主题。为了减少小众主题对实验结果的影响，因此笔者采用C-TF-IDF对主题数量进行压缩。通过迭代计算每个主题之间的余弦相似度，比较主题之间的C-TF-IDF向量，合并最相似的向量，最后重新计算C-TF-IDF向量来更新原有的主题表示，以达到将包含文档数量少的边缘主题与最相似的主题进行合并的目的。最终真菌学的研究主题被缩减为24个，图情报学的研究主题被缩减至24个，哲学的研究主题被缩减为20个。各学科包含文献量最多的前5个主题如表2所示：

表2　重组后的各学科Top5主题

真菌学		图书情报学		哲学	
基因遗传	3 945	医学信息学	7 854	哲学基础理论	24 784
生物物种	3 302	知识组织	7 640	生物与农业哲学	1 498
传染病学	2 577	文献计量	3 205	科学哲学	1 412
临床医学	983	网络信息传播	3 109	医药与疾病哲学	1 383
细胞结构	865	高校图书馆	3 007	健康与护理哲学	758

表2按照各学科主题所含文献数量降序排列。然而，发现3个学科主题分布的特征与差异性并非本研究的目的，此处得出的各学科主题分布特征与主题归纳结果，仅作为后续判识研究人员研究主题转换与迁移的基础。

5 研究结果

5.1 主题相似性与转换概率的宏观分析

为保障实验结果的有效性,研究工作对此前得到的 266 388 位作者进一步筛查。首先删除具有缺失值的数据,其次选取发表文献数不低于 5 篇的科研人员。最终得到真菌学作者 5 427 位,图书情报学作者 3 912 位,哲学作者 1 371 位。将科研人员被引频次最高文献的发表作为判定其达到职业高峰的标准,以科研人员最高被引文献发表当年为职业高峰期(Career Peak,CP),根据公式(3)对科研人员职业高峰期前、职业高峰期当年、职业高峰期后的主题相似度得分进行计算。同时,为了进一步对科研人员研究主题的转换特征进行分析,采用公式(4)对应计算不同时期各学科科研人员主题转换概率。所得结果如表 3 所示:

表 3 高峰期前后主题相似性及转换概率

指标		真菌学	图书情报学	哲学
pre-CP	similarity	0.435	0.600	0.840
	probability	0.271	0.274	0.204
CP	similarity	0.396	0.560	0.849
	probability	0.229	0.186	0.157
post-CP	similarity	0.401	0.584	0.847
	probability	0.304	0.334	0.255

注:pre-CP 表示职业高峰前;CP 表示职业高峰;post-CP 表示职业高峰后;similarity 表示主题相似度;probability 表示主题转换概率

表 3 记录了不同学科科研人员在达到职业高峰之前、职业高峰期当年以及职业高峰之后发表成果的平均主题相似度得分,以及科研人员个体研究主题发生转换的概率。从不同学科科研人员研究主题的整体相似性来看,哲学学科的主题相似度得分最高(≥0.840)。同时,该学科的主题转换概率相比其他两个学科也是最低的。这意味着哲学领域的科研人员在学术生涯的不同阶段内部,个体科研人员不同研究成果的研究主题具有较高的相似性;对应的主题转换概率也表明,哲学领域的科研人员在每个学术生涯阶段内部并不会发生太频繁的主题迁移。另一方面,真菌学领域科研人员研究主题的平均相似度得分最低(≤0.435),说明真菌学领域科研人员在不同阶段内部的研究主题跨度最大;图书情报学领域科研人员在不同阶段的主题相似度得分表明,该领域科研人员的主题跨度略小于真菌学。但是从主题转换概率的计算结果看,图书情报学领域科研人员在高峰期前/后均比真菌学有更大的概率发生主题转换,在高峰期当年主题转换的概率小于真菌学。

上述分析说明不同学科的科研人员在主题相似性与转换概率上存在一定的差异,那么在学科内部从不同时期科研人员平均主题转换概率的计算结果来看,3 个学科科研人员在高峰期之后的研究主题转换概率都要高于职业高峰期之前。真菌学科研人员高峰期之后的主题转换概率相比高峰期之前提高了 12.2%,图书情报学提高了 21.9%,哲学提高了 25%。这一结果表明,科研人员总体队伍在经历了职业高峰期之后,会在不同的研究主题之间更频繁地转换自己的研究方向。当然,也从另一个侧面反映出,科研人在达到职业高峰期之前,在研究主题上具有相对较高的专一程度。在到达职业高峰之前,科研人员更倾向于做自己擅长的、或是这一时期专攻的某项研究;而在职业高峰之后,科研人员开始有更高的职业自由度,不再局限于曾经相对集中的研究主题,因此主题转换的频繁程度会增加。

5.2 精英学者高峰期前后主题转换特征

各学科的精英学者通常是所在学科科技进步的领军力量。在学术界已经关注到精英学者与普通学者在学术生涯与创造力上的差异的同时[33],政府也出台政策加大对科技拔尖人才与优秀科技工作者的鼓励与支持[4]。这部分研究进一步探查精英学者在职业高峰期前后研究主题的转换特征,以期为国家科技政策的制定与实施提供科学依据。目前,学术界对精英学者的识别往往根据其科研成果贡献数量(高发文)、被学术界认可程度(高被引)等指标加以判识。在具体的研究中,兼顾发文量与被引量指标筛选各学科发文数量排名前 1%,且单篇论文平均被引频次排名前 1% 的学者。同时,为使结果具有普遍性,不考虑"一闪即逝"的科研人员,确保从事科学研究不小于 10 年的高发文且高被引科研人员作为领域精英学者展开分析。按照上述标准进行筛选,获得真菌学精英学者 170 位,图书情报学精英学者 246 位,哲学精英学者 97 位。如果说前序的分析关注科研人员在每个时段内的研究主题转换幅度与频繁程度,那么这部分研究则更关注精英学者在以职业高峰为分界线的前后两个阶段的研究主题转换的差异。将每位精英学者在职业高峰期之前和职业高峰期之后所发表的文献分别整合成两个长文档,并分别进行主题建模,然后采用公式(2)计算高峰期前后的主题相似度得分。得到精英学者职业高峰期前后的主题相似度得分分布如图 3 所示,横轴表示学者数量占比,纵轴表示相似度得分区间。

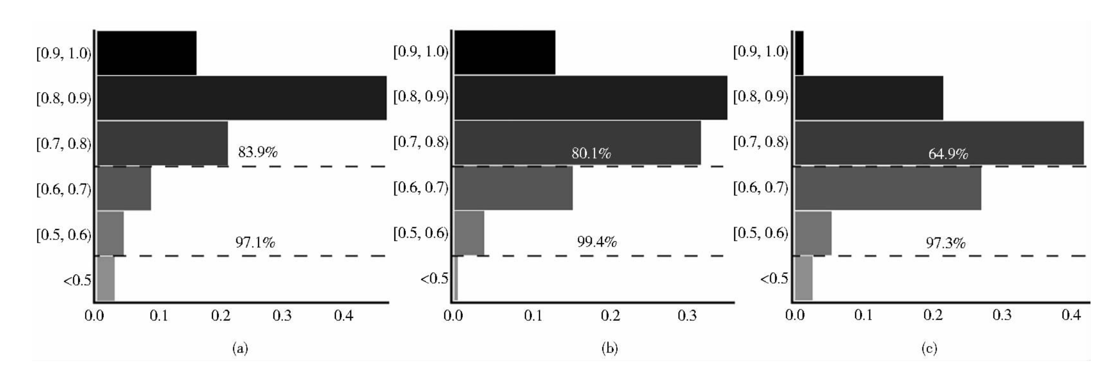


图 3　精英学者职业高峰前后主题相似度

图 3(a)(b)(c)分别为真菌学、图书情报学、哲学领域精英学者在达到职业高峰前后的研究主题相似度得分分布。总体上看,各学科精英学者在经过职业高峰之后,其研究主题与到达高峰期之前的研究主题仍然具有很高的相似性。各学科精英学者在职业高峰期之后所选择的研究主题与高峰期之前相似度在 0.5 以上(≥0.5)的人员占比均达到 97% 以上(97.1%、99.4、97.3%)。在高峰期前后主题相似度大于等于 0.7 的条件下,3 个学科精英学者的占比分别为 83.9%、80.1%、64.9%。各学科精英学者占比最高的主题相似度得分区间分别为[0.8, 0.9)、[0.8, 0.9)、[0.7, 0.8)。这一结果表明,各学科中大多数精英学者都能够在职业高峰期之后仍然保持研究主题的连续性,即使发生一定程度的主题迁移,也依然选择与早期研究非常相近的主题(相似度得分高)。

显然,上述结果与此前针对科研人员总体队伍的分析结果并不完全相符,因此研究工作进一步对各学科精英学者职业高峰期之后与职业高峰期之前主题转换概率的差值进行计算。以横轴表示学者数量占比,纵轴表示差值的区间,精英学者职业高峰期前后主题转换概率的变化情况如图 4 所示:

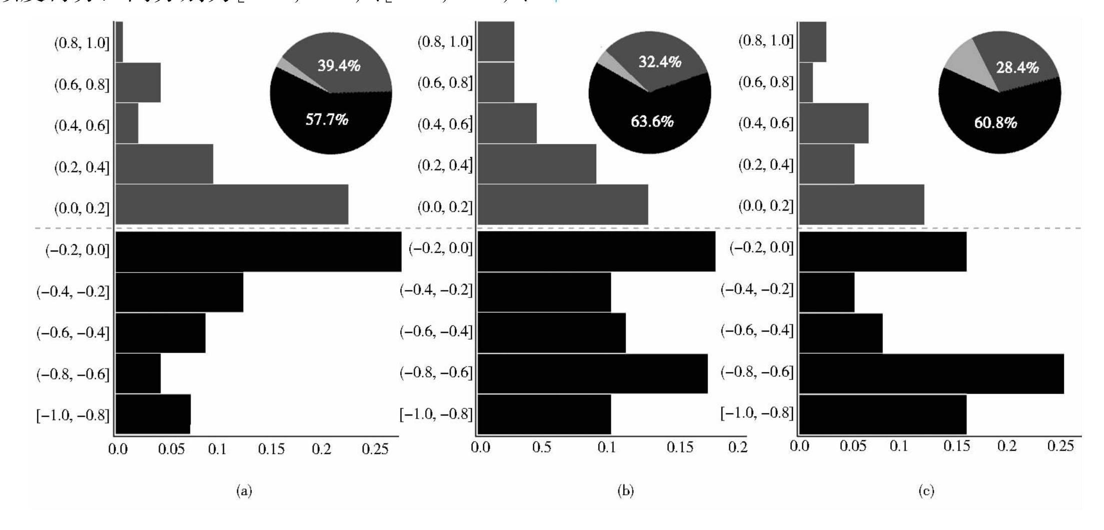


图 4　精英学者职业高峰前后主题转换概率变化

在图 4 中,以虚线为中心越靠近虚线代表精英学者在职业高峰期之后与职业高峰期之前相比其研究主题转换概率变化越小。上半部区域距离虚线越远表示职业高峰期之后主题转换概率增加的值越多(差值靠近 1);下半部区域距离虚线越远则表示职业高峰期之后的主题转换概率降低的值越多(差值靠近 －1)。从图 4(a)(b)(c)反映的真菌学、图书情报学、哲学领域精英学者高峰期前后主题转换概率变化的总体情况看,大多数精英学者在经历职业高峰期之后,其主题转换概率变化的差值处于虚线下方。即这些精英学者在经历个人职业高峰之后,其主题转换概率呈现不变或下降趋势(虚线下方包括差值范围为[－1, 0])。图 4 中内嵌的饼图呈现主题转换概率增加(差值 >0)、降低(差值 <0)、不变(差值 =0)的精英学者占比。从内嵌图的结果看,各学科精英学者在经历职业高峰后,大多数精英学者的主题转换概率相比职业高峰前呈现下

降状态(57.7% >39.4%、63.6% >32.4%、60.8% >28.4%)。其中,哲学领域精英学者表现最为明显,主题转换概率下降的精英学者(60.8%)超过主题转换概率增加的精英学者(28.4%)的两倍。而且,图4(c)中也反映出哲学学科精英学者主题转换概率下降的幅度最大(差值靠近-1的人员占比较多)。这一结果说明,精英学者在经历学术生涯的高峰期之后,倾向于从事比高峰期之前更加专一的科学研究。

6 结论与讨论

笔者采用文献计量学与文档主题建模相结合的方法,对真菌学、图书情报学和哲学3个学科科研人员职业高峰及其相关的研究主题转换特征进行探索。综合上述分析的结果,初步得出以下结论:

(1)科研人员总体上在经历职业高峰之后主题转换会更频繁。在针对科研人员总体的分析中,尽管职业高峰前后的主题相似度差异并不明显,但是主题转换概率这一指标却体现出职业高峰前后的明显差异。各学科的科研人员在经历职业高峰期之后的主题转换率要不同程度地高于职业高峰期之前的主题转换率(参见表3)。这一结果说明就科研人员的总体而言,未达到职业高峰期的科研人员其研究主题转换并不频繁,而经历过职业高峰期之后,科研人员研究主题的转换比高峰期之前更频繁。

(2)精英学者在经历了职业高峰之后其研究主题会更加专一。精英学者高峰期前后的主题相似度表明,大多数精英学者在职业高峰期前后的研究主题具有很高的相似性(参见图3),并且高峰期之后的主题转换概率相比高峰期之前更低(参见图4)。这一结果说明,科研人员中精英学者的主题转换表现出与科研人员总体队伍近乎截然相反的特征:越是在科学研究中表现优秀的精英学者,越在经历职业高峰之后倾向于更加专一的研究方向,其研究主题也越发青睐于"十年磨一剑"。

在科学技术飞速发展的今天,发现和揭示科研人员学术生涯发展过程中的模式与特征,有助于揭示科学生产力发展机制,对于科研管理部门制定积极的科研政策,更好地引导科研人员实现科技创新,具有重要的促进作用。研究中也存在一些不足之处,在自然科学、社会科学、艺术与人文科学中各选择一个学科作为代表,尚不足以覆盖更大范围的科学研究领域。通过主题建模及主题相似度测度科研人员主题转换偏重语义信息,对于更细密的学科与研究方向分类体现尚不完全充分。未来的研究中,将进一步包容更广泛的科学领域,采用更细致的分析方法展开更深入的研究。

参考文献:

[1] 周建中, 闫昊, 孙粒. 我国科研人员职业生涯成长轨迹与影响因素研究[J]. 科研管理, 2019, 40(10): 126-141.

[2] MERTON R K. The matthew effect in science[J]. International journal of dermatology, 1968, 27(3810): 56-63.

[3] LIU L, WANG Y, SINATRA R, et al. Hot streaks in artistic, cultural, and scientific careers[J]. Nature, 2018, 559(7714): 396-399.

[4] 中共中央, 国务院. 关于进一步弘扬科学家精神加强作风和学风建设的意见[EB/OL]. [2021-07-18]. http://www.gov.cn/zhengce/2019-06/11/content_5399239.htm.

[5] RUAN W, HOU H, HU Z. Detecting dynamics of hot topics with alluvial diagrams: a timeline visualization[J]. Journal of data and information science, 2017, 2(3): 37-48.

[6] 邱均平, 余厚强. 科学家黄金年龄影响因素的综合分析[J]. 情报杂志, 2014, 33(3): 11-15, 5.

[7] COLE S. Age and scientific performance[J]. American journal of sociology, 1979, 84(4): 958-977.

[8] JONES B F, WEINBERG B A. Age dynamics in scientific creativity[J]. Proceedings of the national academy of sciences, 2011, 108(47): 18910-18914.

[9] SIMONTON D K. Career landmarks in science: individual differences and interdisciplinary contrasts[J]. Developmental psychology, 1991, 27(1): 119.

[10] SIMONTON D K. Age and outstanding achievement: what do we know after a century of research? [J]. Psychological bulletin, 1988, 104(2): 251.

[11] BRODETSKY S. Newton: scientist and man[J]. Nature, 1942, 150(3816): 698-699.

[12] STEPHAN P, LEVIN S. Age and the Nobel Prize revisited[J]. Scientometrics, 1993, 28(3): 387-399.

[13] LI J, YIN Y, FORTUNATO S, et al. Scientific elite revisited: patterns of productivity, collaboration, authorship and impact[J]. Journal of the royal society interface, 2020, 17(165): 20200135.

[14] JONES B F. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? [J]. The review of economic studies, 2009, 76(1): 283-317.

[15] COKOL M, IOSSIFOV I, WEINREB C, et al. Emergent behavior of growing knowledge about molecular interactions[J]. Nature biotechnology, 2005, 23(10): 1243-1247.

[16] SINATRA R, DEVILLE P, SZELL M, et al. A century of physics[J]. Nature physics, 2015, 11(10): 791-796.

[17] PETERSEN A M, FORTUNATO S, PAN R K, et al. Reputation and impact in academic careers[J]. Proceedings of the national academy of sciences, 2014, 111(43): 15316-15321.

[18] PETERSEN A M. Quantifying the impact of weak, strong, and su-

per ties in scientific careers[J]. Proceedings of the national academy of sciences, 2015, 112(34): e4671－e4680.

[19] 史庆伟，乔晓东，徐硕，等. 作者主题演化模型及其在研究兴趣演化分析中的应用[J]. 情报学报，2013，32(9)：912－919.

[20] 陈立雪，郭思月，滕广青，等. 科研人员研究主题的聚焦与迁移研究[J]. 数字图书馆论坛，2019(12)：9－17.

[21] UZZI B, MUKHERJEE S, STRINGER M, et al. Atypical combinations and scientific impact[J]. Science, 2013, 342(6157): 468－472.

[22] GUIMERA R, UZZI B, SPIRO J, et al. Team assembly mechanisms determine collaboration network structure and team performance[J]. Science, 2005, 308(5722): 697－702.

[23] BOURDIEU P. The specificity of the scientific field and the social conditions of the progress of reason[J]. Social science information, 1975, 14(6): 19－47.

[24] HOONLOR A, SZYMANSKI B K, ZAKI M J. Trends in computer science research[J]. Communications of the ACM, 2013, 56(10): 74－83.

[25] RZHETSKY A, FOSTER J G, FOSTER I T, et al. Choosing experiments to accelerate collective discovery[J]. Proceedings of the national academy of sciences, 2015, 112(47): 14569－14574.

[26] JIA T, WANG D, SZYMANSKI B K. Quantifying patterns of research-interest evolution[J]. Nature human behaviour, 2017, 1(4): 78.

[27] ZENG A, SHEN Z, ZHOU J, et al. Increasing trend of scientists to switch between topics[J]. Nature communications, 2019, 10(1): 1－11.

[28] HOFMANN T. Probabilistic latent semantic indexing[C]//Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1999: 50－57.

[29] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 3(1): 993－1022.

[30] ANGELOV D. Top2Vec: distributed representations of topics[EB/OL]. [2021－02－18]. https://arxiv.org/pdf/2008.09470.pdf.

[31] SALTON G, YU C T. On the construction of effective vocabularies for information retrieval[J]. Acm sigplan notices, 1973, 10(1): 48－60.

[32] SINATRA R, WANG D, DEVILLE P, et al. Quantifying the evolution of individual scientific impact[J]. Science, 2016, 354(6312): 596.

[33] FROSCH K H. Workforce age and innovation: a literature survey[J]. International journal of management reviews, 2011, 13(4): 414－430.

作者贡献说明：

陈立雪：数据采集与分析，论文撰写；

滕广青：提出研究思路，设计研究方案，论文撰写与修订；

吕晶：数据分析；

庹锐：数据分析。

Identification of Characteristics of Topic Change before and after Career Peak of Scientists

Chen Lixue　Teng Guangqing　Lü Jing　Tuo Rui

School of Information Science and Technology, Northeast Normal University, Changchun 130117

Abstract: **[Purpose/significance]** Exploring the individual career development of scientists and the transforming laws of research topics can not only reveal the internal mechanism of the development of scientific productivity, but also help provide better policy guidance and support for the development of scientific undertakings. **[Method/process]** Based on the representative discipline data of natural sciences, social science, art and humanities, this article identified the career peaks of scientists. The career peak was used as the basis for dividing the academic career of scientists. The Top2Vec topic modeling method in natural language processing was used to identify research topics, and the topic similarity and topic transfer probability of the research topics at different stages of the academic career of scientists were measured. **[Result/conclusion]** The research results show that scientists in various disciplines generally change research topics more frequently after experiencing their career peaks, while elite scholars have more specific research topics after experiencing their career peaks.

Keywords: scientists　career peak　top2vec　topic change　topic similarity